Summer 8-31-2002

An algorithm for estimating the quality of microarrays

Ajeet S. Sodhi
New Jersey Institute of Technology

ABSTRACT

AN ALGORITHM FOR ESTIMATING THE QUALITY OF MICROARRAYS

by
Ajeet Sodhi

Microarray technology is currently one of the most valuable gene expression tools in molecular biology allowing the experimenter to simultaneously quantify the expression of thousands of genes. It is also one of the most difficult tools to use accurately as each microarray produces a large amount of information that needs to be inspected and normalized before analysis. As the size of a microarray or number of replicates increase, the use of manual inspection becomes impractical. The aim of this thesis is to introduce an algorithm that evaluates each feature of a microarray from the scanned data file. A quality score is calculated from various spot parameters, the *quality-quotient*, and can be used to automatically assess the quality of the spot. This *quality-quotient* can then be utilized to automatically select quality spots or act as a weighting factor for comparing spots from replicate microarrays.

AN ALGORITHM FOR ESTIMATING THE QUALITY OF MICROARRAYS

by
Ajeet Sodhi

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Federated Biological Sciences Department

August 2002

APPROVAL PAGE

AN ALGORITHM FOR ESTIMATING THE QUALITY OF MICROARRAYS

Ajeet Sodhi

Dr. Michael Recce, Thesis Advisor — Date
Director, Life Sciences Program and Center for Computational Biology, NJIT

Dr. Ronald P. Hart, Committee Member — Date
Professor of Biology, Rutgers University

Dr. Peter Tolias, Committee Member — Date
Director, Center for Applied Genomics
Public Health Research Institute

BIOGRAPHICAL SKETCH

Author: Ajeet Sodhi

Degree: Master of Science

Date: August, 2002

Undergraduate and Graduate Education:

- Master of Science in Computational Biology,
  New Jersey Institute of Technology, Newark, NJ, 2002

- Bachelor of Science in Biology,
  College of William and Mary, Williamsburg, VA, 2000

Major: Computational Biology

ACKNOWLEDGMENTS

I would like to thank Dr. Ronald Hart and Dr. Michael Recce for their understanding, for their patience, and for all their help and support. I would not have been able to finish this thesis without them. I would also like to thank (in alphabetical order) Jason Carmel, David Kahn, Jessica Lam, and Jonathan Pan for their valuable assistance as the "experts" and for their insight during brainstorming sessions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Functional Genomics is the study of all of the genes in a cell at the transcriptional level. This refers to the measuring of the concentrations of mRNA of a sample. Techniques involved include Quantitative Real-Time PCR, Serial Analysis of Gene Expression, and Northern Blotting, all of which vary with respect to the precision, accuracy, and number of genes that can be simultaneously profiled. For sheer quantity of genes assayed however, few methods compare to the use of nucleic acid microarrays at this time. This technique is currently being employed for purposes ranging from clinical diagnosis [17] to experimental biology in areas as diverse as spinal cord injury research [23, 24] and plant genetics [18].

Although the use of nucleic acids immobilized on solid surfaces for the purposes of quantifying biomolecules has been a core technique of molecular biology for decades, high density microarrays are a relatively recent tool and allow profiling of the expression of thousands of genes in parallel [1, 2, 3, 4]. Microarrays are composed of slide surfaces upon which numerous spots have been printed by a robot. The spots are *probes* consisting of nucleic acids representing genes obtained from a gene library and are affixed to the surface during the printing process. Each spot represents a different gene.

Microarrays may contain probes consisting of either cDNA or oligonucleotides [1]. cDNA is the artificial nucleic acid formed from the transcript, mRNA. This is accomplished with the enzyme *reverse transcriptase*, which produces DNA from RNA, a reversal of the normal process. Because of this, the cDNA is exactly complementary to its mRNA precursor. It is also much shorter than the original gene copy that the mRNA was derived from since the production of mRNA eliminates the non-coding DNA regions known as

introns. The result is a nucleotide polymer that can be used as a probe for a specific gene transcript.

In a typical microarray experiment, the tissue or cell specimen that is to be tested is isolated and processed, during which the total mRNA is extracted from it. This sample contains a representation of all of the genes that were being expressed in the cells or tissue at the time of harvesting. The sample is reverse-transcribed to cDNA along with the addition of nucleotide fluorophores. This fluorescently labeled cDNA strand is referred to as the *target* and is hybridized to the microarray.

Theoretically, the cDNA targets will bind to their specific complementary cDNA probes. During scanning, a laser excites the labeled probe-target hybrids, inducing them to fluoresce. The intensity of the fluorescent emissions is recorded and is indicative of the quantity of bound target [4]. Additionally, the intensity of the emissions for a spot is directly proportional to the concentration of transcript in the original sample.

Usually, a microarray experiment is performed using two samples, a control and a test sample. This is the most robust method to determine the change in gene expression between two treatments. The samples are labeled with different fluorophores, Cyanine 3 (Cy3) and Cyanine 5 (Cy5) and hybridized to the microarray slide simultaneously. Two lasers are used during scanning, each designed to excite only one of the fluorescent labels. Each laser is employed separately. The resulting emissions are combined computationally and give a simple, visual report of the findings.

Software, such as Genepix 4.0 (Axon Instruments Inc.), is used to analyze the results of a microarray scan. This program produces a composite image of the information resulting from the emissions of both of the fluorophores. A red spot indicates that the red fluorophore

(Cy 5) is primarily found within the spot. Thus, the targets bound to the probes within the spot are labeled with Cy 5, meaning that the Cy5-labeled sample is showing expression for the given gene. Conversely, a green spot results from the binding of the targets from the Cy 3-labeled sample to the probes within the spot. The intensity of the red and green colors refer to the intensity of their respective emissions. Finally, spots containing targets from both samples are shown in an artificial yellow color [Fig. 1].

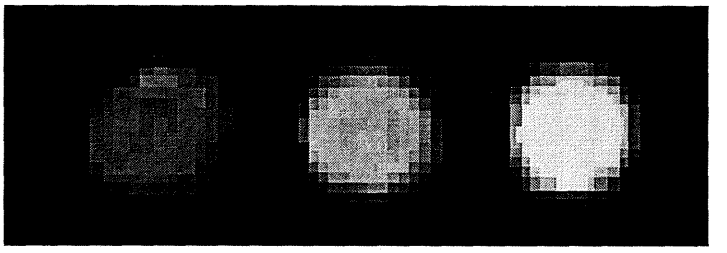

Figure 1 Spots resulting from bound Cye 5, Cye 3, and the mix of the two, respectively.

The microarrays discussed in this thesis use a modification of the standard microarray procedure. The probes are printed with oligonucleotides instead of cDNA. Additionally, the targets are modified to incorporate fluorophores through the use of dendrimers.

The microarrays are rat oligonucleotide microarrays and are used to study the change in gene expression of rats following spinal cord injury. Each microarray consists of 4969 spots, with probe lengths of 65-70 nucleotides [6]. The use of short oligonucleotide probes was introduced to reduce the problem of non-specific hybridization. The extensive redundancy in the genomes of organisms, especially eukaryotes, allows targets to hybridize to unintended probes with a frequency severe enough to render unreliable results. Short oligonucleotide probes, usually about 50-70 nucleotides long, can be intentionally designed

to be long enough for specific target identification, but short enough to reduce the redundancy contributing to non-specific hybridization [1, 5].

The cDNA targets are modified to utilize dendrimers. The advantages of this technique include a higher signal to background ratio and a decrease in the initial mRNA required [21]. Following processing and isolation, the total RNA pool is reverse-transcribed using an oligo-dT primer with a dendrimer "capture" sequence attached. The fact that nearly all mRNA transcripts end with a repeating poly-adenosine (poly-A) nucleotide sequence ensures that they will all be reverse-transcribed. Afterwards, heat and alkali substances degrade the RNA leaving only the reverse transcription product. The result is a pool of cDNA representative of the original total mRNA transcripts, each with an identical capture sequence attached. The process is repeated for the second sample except a different capture sequence is used.

After hybridization with the two samples, the slide is incubated with the dendrimers, large molecules of branching oligonucleotides containing fluorescent dyes (Cy 3 or Cy 5). Two types of dendrimer are used, each designed for one capture sequence and each containing either Cy3 or Cy5. The dendrimers bind to their specified capture sequences and fluoresce during scanning.

It is necessary to define some terms that will be used in this discussion. The parameters associated with the *feature* of a particular gene refer to the actual spot on the microarray. The *background* is a region surrounding a particular feature. It is individually defined for every spot on the microarray. The inner boundary of the background is defined as two pixels from the outer edge of the feature. The outer boundary is a circle extending around the current feature for a radius that is three times the radius of the current feature,

excluding the boundaries of the surrounding spots (Fig. 2). All background calculations for this spot are derived from within this bounded region.

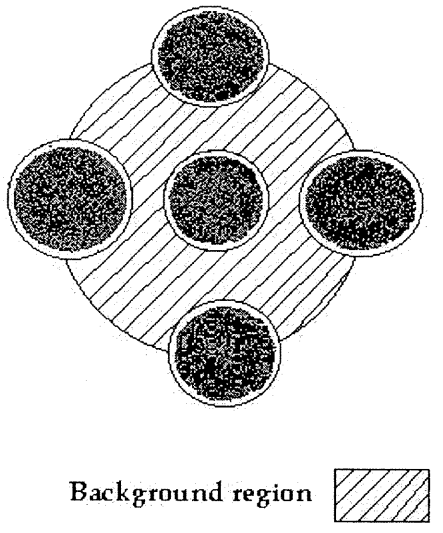

Figure 2 The striped region denotes the background for the central spot.

Although the theoretical foundations of microarrays are straightforward, in practice there are numerous conditions that may lead to substantial systematic variation. The result is that normalization procedures must be applied to raw microarray data before reliable expression data may be extracted. Sources of variation include dye biases dependent upon intensity or spatial location within the slide, and variation due to minuscule differences in printing hardware.

The second source of variation, known as print-tip variation is one of the most important considerations when analyzing high-density microarrays. The probes on a microarray are printed in discrete subunits. The current microarray configuration consists of 32 blocks of probes, 4 across and 8 down (Fig. 3). Each block consists of 168 probes, 12 across and 14 down. During the printing process, the probes within an individual block are

all placed upon the slide by the same printing pin. The pin resembles a slotted metal rod, which draws up the probe in liquid form by capillary action and releases it onto the proper location on the slide upon contact. Minuscule differences in pin structure produces significant differences in probe capacity from block to block [22].

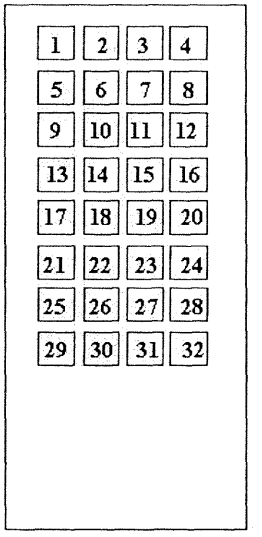

Figure 3 Block layout on microarray slide.

Print-tip normalization is accomplished using the Statistical Microarray Analysis function [20] in 'R', a statistical software package based on the commercial S+ software [19]. The function transforms data for a single microarray to be displayed as a log of intensity ratio vs. log of mean intensity and then finds a best fit to the scatter of points [22].

CHAPTER 2

THE PROBLEM

The important advantage of the microarray is the ability to assay thousands of genes concurrently. And while the information obtained from the microarray does show general trends in gene expression, its accuracy does not compare with other, more sensitive methods. In one example, a two-fold increase in expression levels from a microarray was revealed as a 23-fold increase by the more sensitive method of Quantitative Real-Time PCR [16]. For this reason, microarrays are used as a primary screening process, by selecting promising looking genes for further study by more accurate methods.

The enormous amount of information generated by a microarray experiment presents a problem for the researcher. Even a relatively small microarray with only a few hundred to thousand spots produces overwhelming spreadsheets of data. The researcher can only superficially examine the chip scan under a program such as Genepix to eliminate obvious locations of error before submitting the image to normalization and analysis. It is not practical to devote the necessary time and energy required to thoroughly examine a scan manually, particularly with the necessary replicate arrays. Therefore, numerous discrepancies are neglected that may still contribute to the intensity scores, possibly influencing the data.

The result is that gene expression data may be obtained that appear to be highly significant, but may have been influenced by errors within the chip. These types of errors need not be a disaster for the entire experiment. Even a well-performed microarray chip may have a significant population of questionable results. Therefore, a method that can automatically discriminate between doubtful and reliable spots would be very useful.

The purpose of this thesis is to introduce a quality-control function that can be used to inform the researcher as to the reliability of his gene expression data. This function is applied to the microarray data before the normalization procedures. The function calculates a single score, a *quality-quotient* or *qq* for every spot within the raw data file. The *qq* is a composite of several factors that examine the different aspects of a spot on a microarray. The range of the *qq* is defined to be between 0 and 1. A researcher producing a *qq* score close to 1 for a particular gene would know that the values obtained for the spot was probably authentic and not an artifact of experimental error. Additionally, the distribution of *qq* values can be analyzed for the entire chip to give an idea of the overall chip quality.

Other assessments of microarray quality exist, but vary according to usefulness or application. The most common procedure is to have methods designed to assess variability placed into a chip. Multiple spots of the same gene on a single chip is the obvious design. The values for a given spot can immediately be compared to the values of its sister spots revealing the extent of any within-chip variation [10, 11, 12]. While useful, this technique is not practical because of space limitations on microarrays. Similar across-chip assessments also suffer from the same weakness.

The use of the actual scanned microarray image is the ideal source of information for any quality evaluating procedure. The image is the primary receptacle from which the expression information is derived. It holds a pixel-by-pixel account of the intensities generated by the laser stimulation. As such, it holds the largest amount of information from which the most reliable data may be extracted. Most of the quality evaluation papers have focused on this source [11, 12, 13].

Matarray, a quality-assessment program, was demonstrated by Wang, Ghosh, and Guo [13]. Here, intermediate scores are calculated from the microarray image scans to record irregularities in spot intensity, size, and background noise levels. From these, an assessment of individual spot quality is produced for every spot. This program, and others like it, requires a fully descriptive, information-laden microarray image for the algorithm to run on. The algorithm must examine the individual pixels that comprise the image.

The data file produced from the primary image by Genepix is a text file and contains reduced information for every spot. Descriptive calculations have been tabulated for each gene spot as a whole [8]. It was thought that if a quality control function could be implemented directly from the information contained in the text file, the function could be activated without the use of the actual image and be implemented at the same time as the automation of the normalization and cleaning functions. It is noted that any assessments made from the secondary text file would tend to not be as accurate as the functions directly utilizing the primary image because of loss of information. However, the use of a quick, simple quality predictor does carry advantages in the experimental field.

In order to produce the quality estimating functions that make up the quality-quotient, it was necessary to analyze the results of microarray experiments, including both the scans and the normalized intensity values. Usually, the normalization procedure is performed manually in 'R'. The 'R' interface requires numerous lines of instruction for each microarray slide. As well as being time consuming, this portion of the normalization protocol presents frequent opportunities for error. In order to increase the efficiency of this section of the normalization procedure, a function was written in 'R' to automate it (Appendix A).

The purpose of this thesis is to describe the methods used to conceive the various aspects of the quality control function. Following this, the function was tested on a set of microarray spots. The results were compared to scores for the same set of spots evaluated by individuals skilled in microarray analysis.

CHAPTER 3

METHOD

The parameters that compose the final quality control factor, or *qq*, were selected by careful inspection of the data from numerous microarray experiments. It is worth noting that none of these parameters, as individuals, are excellent predictors of spot quality alone. This is to be expected, as they each measure a different aspect of the microarray spots. There are many factors that may detract from spot quality. It is only with a composite score that a realistic, overall assessment of a spot may be obtained.

After scanning, a microarray image is analyzed by the microarray analysis program Genepix. Following analysis, the intensity values as well as several parameters calculated by Genepix to aid analysis and normalization are placed in a file with a ".gpr" extension. This data file is the source for the parameters discussed below.

3.1 Feature-Background Ratio

The first parameter was chosen to be a simple ratio expressing the extent to which the feature fluorescence surpassed the background fluorescence for a particular gene and wavelength. This was calculated using the feature median and background median. These two variables are also the usual subjects of the normalization and analysis procedures since they are the least sensitive to the random noise of the system.

The formula for the first parameter, feature-background ratio (FB) is,

$$FB = \frac{F_{med}}{B_{med}}$$

where F stands for feature, B stands for background, and med is median.

The analysis manual *Biological Relevance of Genepix Results* [7] suggests a minimum FB ratio of 2, or having the feature value at least twice that of the background. Experience has shown this to be a conservative score and that there are results that are significant with a lesser FB, particularly when other feature qualities are assessed simultaneously.

Additionally, there have been recent improvements to the experiment protocol, reducing noise and creating a more sensitive system. The hybridization step of the experiments were originally done manually, by placing both fluorescently labeled cDNA samples on the surface of the chip in an amount of liquid and allowing the setup to sit immobile overnight. With the introduction of the Ventana Discovery (Ventana Medical Systems), a hybridization-automation machine, the target samples are forced to circulate on the chip surface in a larger amount of liquid by the jet-blowing action of the machine. The result is a dramatic drop in non-specific hybridization. The backgrounds images of the microarrays hybridized by the Ventana are notably more uniform resulting in an overall noise reduction in the system.

The combination of experience and the cleaner images has led to the conclusion that an acceptable level of the FB ratio is '1.4'. In order to incorporate the raw FB score into the overall quality assessment, it needs to be transformed with a scoring function into an intermediate score with a value between 0 and 1.

A percent ranking function was considered. This function assigns a percent score to a member of an array according to the relative rank of the values with the other members of the array. While this method did fulfill the objective of assigning a meaningful score to the raw value, it was eventually rejected as being too relative. The transformed score should be

independent of its relative status in the array data file. This would allow quality scores from a particular array to be compared with scores from a different array. A relative ranking method is capable of giving the same score for two different arrays even though their respective qualities differed significantly.

Several other functions that would transform the raw score to a number between 0 and 1 were considered. The "squashing" function that was chosen is part of the arctan function:

$$\text{FB score} = (2/\Pi) \arctan 2(\text{FB}-1.4)$$

In addition to the ability to convert the raw score to a number between 0 and 1, the high slope at the start of the function allows values that begin at the threshold to immediately receive good scores, instead of being penalized by their proximity to the cutoff as in a sigmoidal curve (Fig. 4).

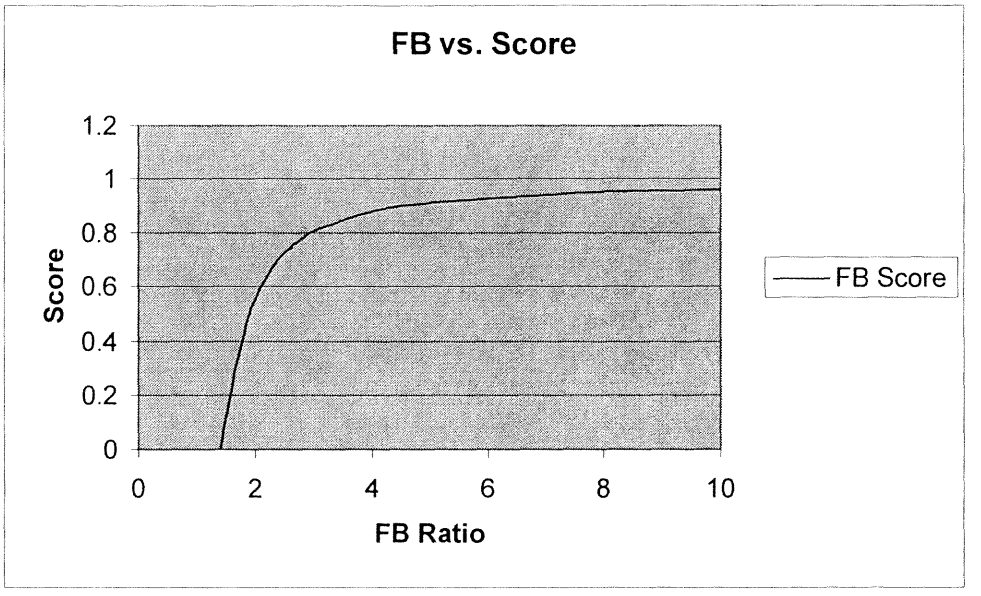

Figure 4 Relationship of FB score vs. FB ratio.

It was desired that the next component of the overall quality score incorporate different measured characteristics of the microarray datasets. By including unrelated aspects, it was hoped to get as complete an assessment as possible of the data. The next several parameters all take into account the means and variances each spot.

3.2 Z-Score

The next parameter looked at was a calculation resembling a z-score obtained for a single gene at both the red and green wavelengths:

$$\frac{F_{\mu} - B_{\mu}}{B_{\delta}}$$

This calculation was also reviewed with the median used in place of the mean. When graphed as a function of the mean, it showed a general positive correlation with the increase of the mean. This is required because it would be expected for the quality of a spot to increase proportionally in this manner. As the measurement of a feature extended further away from the background levels, it would become easier to distinguish the feature from the general noise of the system.

Despite the positive trend of this parameter, it was ultimately rejected as a measure of spot quality. The central use of the standard deviation of the background within the calculation rendered it exceedingly sensitive to fluctuations caused by the imperfections in the microarray design.

During the hybridization step of a microarray experiment, the two target samples labeled with different fluorophores are placed on the surface of the microarray chip for a length of time to allow the transcript targets to anneal to their complementary probes. Although the majority of transcripts arrive at their destination, non-specific binding does

occur. There are many potential sources of this non-specific hybridization. There may be imperfections in the slide glass or surface coating that non-specifically trap the molecules. There are probably certain gene targets for which no probe on the chip exists. There might even be minute drops of probes unintentionally splattered by the robot during the manufacturing of the chip. The scanned images of the microarrays never fail to show bright pixels of both fluorophores resting in areas on the chip that are designated as background, even in a well performed experiment.

The results of these imperfections are occasional disruptions of the calculations. The intensity of a bright pixel or smudge caused by non-specific hybridization can render a background standard deviation calculation many times above the actual value, making it useless for this purpose (Fig. 5). Since this problem occurs frequently (an average of 1/12 spots) and since it also did not accurately assess the quality of a spot, it was rejected as a quality measure.

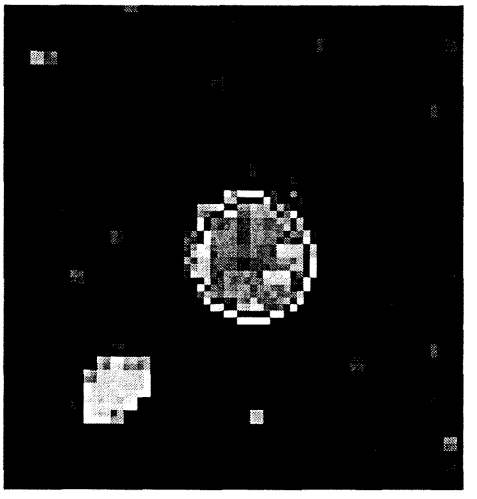

Figure 5 Background standard deviation adversely affected by artifact in lower left.

3.3 Coefficient of Variation

In an ideal experiment, the pixels of a feature would be expected to show a uniform intensity with no variation. Every unit on the entire slide would be expected to show an equal distribution of fluorophores. This is certainly not possible. There are numerous sources of variation, even in the most well performed experiment. The result is that a pixel-by-pixel analysis of any spot presents a distribution of intensity. A large distribution of pixels may be due to dust on the slide, surface irregularities of the slide, anomalies arising from printing such as cross-contamination or pin deformation, or many other contributing sources.

Therefore, it is known that the quality of a spot on a microarray is inversely proportional to the intensity variance [12, 13]. The purpose of the next parameter, the Coefficient of Variation (CV), is to take the variation in pixel intensity into account. The CV is obtained with the following formula:

$$CV = \frac{\text{SD of F}}{F_{med}}$$

SD = spot Standard Deviation, F = spot feature, and med referring to the spot median. Unlike the previous measure, the CV is not as sensitive to the occurrence of random non-specific hybridization or experimental artifacts such as dust. The differing characteristic is the use of the feature SD instead of the background SD.

The low background intensity levels are easily disrupted by the random appearance of the high-intensity pixels caused by non-specific hybridization. The extreme contrasts are manifested in the abnormally high SD of the backgrounds, rendering them useless for inclusion in calculations as discussed before.

A feature usually presents a higher overall degree of fluorescence than background. This allows it to absorb a similar quantity of non-specific hybridization without significantly

affecting the spot measurements such as median, mean and SD, up to a certain point. As the number of intense, non-specific pixels increase, the feature begins to lose the benefits of this buffer zone. The presence of the disturbances begins to affect the other characteristics of the feature.

To present an intermediate CV score for the overall quality score, the result of the CV is transformed by a different squashing function (Fig. 6):

$$\text{CV Score} = 1 - \frac{1.2}{1.2 + 0.07(CV^{-5})}$$

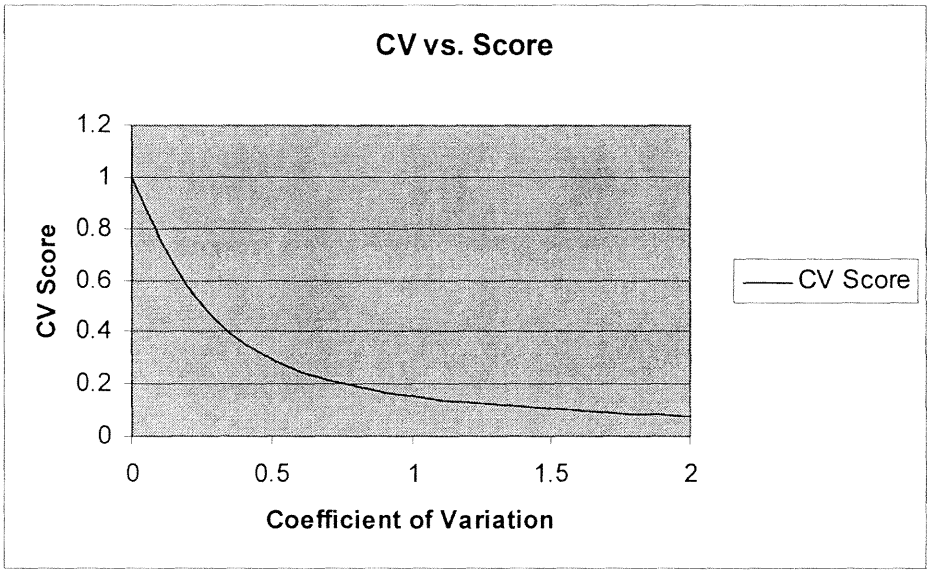

Figure 6 Graph of CV vs. CV score.

The CV scores for both the red and green wavelengths of a spot are averaged before inclusion into the overall composite score.

3.4 Difference of Ratios

The Genepix manual, *Biological Relevance of Genepix Results* [7] describes three calculations derived from a scanned image that are useful to assess spot quality. These calculations attempt to describe the same measurement, specifically the ratio of wavelength intensity. These are the Ratio of Medians (rM), Median of Ratios (Mr), and the Regression Ratio (rR). Each calculation is performed by different methods. A good quality spot will produce similar values among the three calculations. They are calculated for every spot in the microarray. The use of the median is preferred, as before, since the median is less sensitive to outliers.

The rM is a "whole" feature calculation. After scanning, every pixel of a given spot has values for the red and green intensities. Assuming that the ratio consists of red/green in this discussion, the median of the red intensities is divided by the median of the green intensities, for the entire spot, after correcting for the background. Pixels located in the boundary are not included.

Unlike the rM, the Mr (Median of Ratios) is derived by calculating the red/green ratio for every pixel (again subtracting background), and then taking the overall median.

The rR (Regression Ratio) is an independent ratio calculated using the all pixels associated with a spot including feature, background, and boundary. A regression line is defined based on the histogram of all spot associated with a feature. The slope is the rR.

The Genepix manual states that the three ratios, rM, Mr, and rR can be indicative of abnormalities as these three values become more dissimilar for a given spot [8]. High quality microarrays will produce ratios of rM, Mr, and rR that are very similar. Large discrepancies between the values can indicate problems in the experimental procedure.

The third parameter is a measure of the difference among the ratios. For a given spot, the three ratios are averaged to provide a reference. Each ratio is then subtracted from the reference. The absolute values of the differences are added and divided by three. This is now the Mean Deviation (MD) of the ratios:

$$MD = \frac{\sum_{i=1}^{n=3} |\bar{X} - X_n|}{3}$$

In order to normalize to other values, MD is divided by the reference value resulting in the Percent Difference (PD). This is the value describing the overall difference in values amongst the three ratios for a given spot.

In an ideal spot, there would be no difference among the three ratios. Therefore, the highest score belongs to a PD of zero and descends from there. The scoring function of the PD is (Fig. 7):

$$\text{PD Score} = 1 - \frac{2.1}{2.1 + 0.12(PD^{3.1})}$$

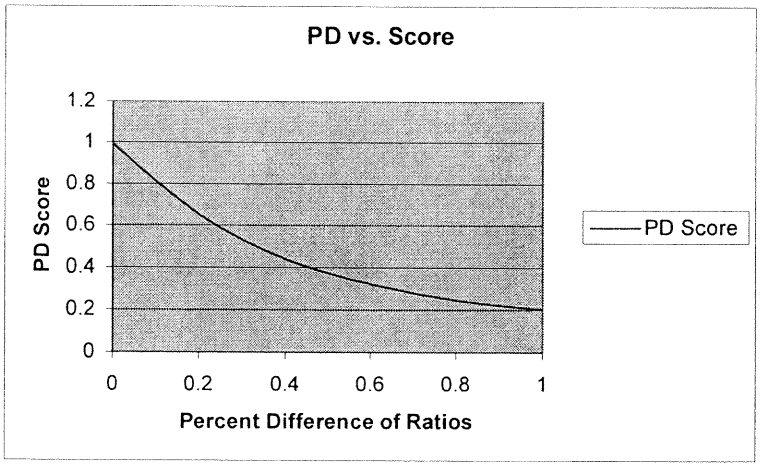

Figure 7 Graph of PD score.

3.5 T-Test

Other measures were tested. It was thought that a statistical type of measure might give a promising perception of the quality of a spot on a microarray. The obvious choice is a t-test calculation, which incorporates the means and the variances of two distributions. The result of a t-test is a decimal describing the probability that the two distributions are significantly different. In this case, the two distributions would be the feature intensity and the background intensity and the aim of the test would be to describe the extent to which the feature could be distinguished from the background.

T-tests are used to analyze microarray data, but in a different respect than here. Currently, the t-test is being used to distinguish results among treatments using multiple spots on the same chip [8], a more familiar capacity.

To utilize the t-test, *t* values were calculated for both wavelengths between every feature and background using only the means, standard deviations, and populations (pixel number) as demonstrated in *Intuitive Biostatistics* (Motulsky, 1995) [8]. The *p* values were obtained using the *t* values and the appropriate degrees of freedom. The resulting probability values were examined as potential contributors to the quality analysis procedure.

There were several notable problems with the results of the t-test as a measure of individual spot quality. Students' t-test is designed to analyze the difference between means with known variances of datasets with small degrees of freedom (small samples of $n<30$) [9]. As such, it is enormously useful in biology experiments where it is used to inquire if the difference between control and test sample sets as small as $n=3$ are significant. As the number of samples increase, it becomes easier to distinguish their distributions leading to more significant probabilities (lower probabilities that the two samples are actually the same). Because every spot on a microarray contains an average of 700 total pixels (including

background pixels), the probabilities produced on the t-test had ridiculously low values, producing underflow errors on more than half of the spots.

Eventually, the t-test was rejected as a measure of quality. Besides the difficulty of producing reasonable, meaningful measurements, the task of evaluating the extent to which the feature could be adequately distinguished from the background was adequately handled by the Feature-Background Ratio.

The formulas for the Quality-Quotient are listed in Appendix B. At this time, the function is performed within Excel. It was considered to code the function, but that would restrict the ability to improve it. The Excel implementation was found to give the most dynamic view of the functions as the formulas and thresholds can be manipulated and observed most easily in the spreadsheet format.

CHAPTER 4

RESULTS

The quality control algorithm was subjected to a comparison test to gauge the accuracy of its predictions. Three "experts" were recruited to independently evaluate a total of sixty spots from three separate chips in a double-blind quality assessment test. The experts were individuals who work with microarrays on a daily basis, in all aspects of microarray experiments, including preparation, RNA extraction and labeling, hybridization, washing, scanning, normalizing, and analysis. The spots were chosen to represent a range of qualities, intensities, and background noise. Each expert was instructed to manually inspect every spot on the list of spots in order, by whichever methods they were accustomed to. This including examining the spots at different magnification rates, constructing histograms of various parameters, and viewing through separate wavelengths.

A simple grading index was given to apply to each spot:

5 – excellent spot

4 – good or acceptable spot

3 – average spot

2 – probably unreliable, questionable spot

1 – unquestionable bad spot

Afterwards, each expert's score was averaged and placed alongside the *qq* score calculated by the quality control function (Appendix C). Figure 8 displays five spots and their corresponding expert average.

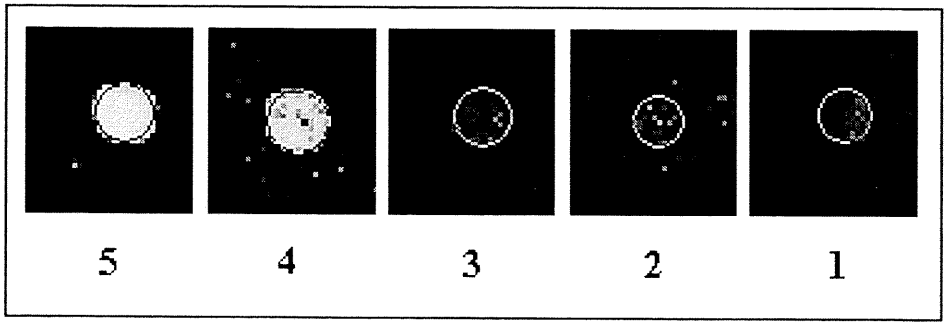

Figure 8 Range of expert scores.

It was first decided to estimate the ability of the *qq* to replicate 'expert' scores. The *qq* score was multiplied by 5 to bring it to the scale of the expert scores (ES). Although this procedure is not entirely accurate because the ES ranges from 1-5 and the *qq* ranges from 0-1, it was deemed sufficient for the purposes of evaluation. A correlation test revealed a value of 0.49. This is not surprising however, since the *qq* is continuous and the ES is an index.

Another thought was to measure the average difference between the *qq*. The ES average was compared to the modified *qq* score. The mean difference between them was calculated for the entire set. The result was a mean difference of 0.71, or less than a quality point of difference on average.

The ability to select quality spots was tested next. For accuracy, it was decided to leave the ES and *qq* their respective original scales. Statistical tests were not included since it was thought that they would produce unreliable information since there is a large amount of subjectivity involved in the judgment of an "average" score. This can be shown by the disparity seen in the experts' selection of average scores. This problem, however, is not too severe for the task at hand. The entire aim of the algorithm is to bring good spots to the attention of the researcher without the time and attention it would require manually.

Thus, the comparison is conducted by the ability of the algorithm to correctly select spots that will provide useful information. In the table, any spot with an expert average of 4 or 5 (good spots) are treated as desirable (Table 1). By setting a threshold of 0.75 for the *qq*, it can be seen that the function correctly selects 9/11 of these good spots, or 82% of them. However, in addition to correctly selecting these 9 spots, it also selects 10 false positives, an error rate of 20%. By raising the *qq* threshold to 0.80, the detection of the positive spots selected by the expert's dips slightly to 8/11, or 73%. This is offset by the decrease in false positive rate to 7 spots, or 14%. By increasing the *qq* to 0.85, the detection rates stays stable at 73%, but the false positive rate is lowered to 3/49, or just 7.6%. Thus, the *qq* is capable of imitating an expert's selection of quality spots most effectively at a threshold of 0.85.

Table 1 Threshold *QQ* Ability to Select True and False Positives.

Threshold	Positive (11 total)		False Positive (49 total)	
QQ	#	percent	#	percent
0.5	10	91%	44	90%
0.55	9	82%	37	76%
0.6	9	82%	31	63%
0.65	9	82%	24	49%
0.7	9	82%	19	39%
0.75	9	82%	10	20%
0.8	8	73%	7	14%
0.85	8	73%	4	8%
0.9	7	64%	2	4%

CHAPTER 5

CONCLUSIONS

In the last few years, the pace of biological research has increased to a blinding speed. Gone are the days when whole laboratories devoted years to the study of the expression of a single gene. It is now known that all gene expression is interrelated and that to study them requires a more global approach. Such approaches are available today with tools like nucleic acid microarrays.

Unfortunately, microarrays produce overwhelming bottlenecks of data that require inspection, normalization, and analysis before it can be of any use. Only computers can efficiently handle these types of tasks. The existence of a quality control procedure that works concurrently with microarray analysis would greatly aid the researcher, since the reliability of the information from microarray experiments is not readily obvious.

In this paper, the creation of such a quality control procedure has been outlined. It has been shown how the careful analysis of raw microarray datasets were examined to produce three measures, the Feature-Background Ratio, the Coefficient of Variation, and the Percent Difference of Ratios and how the scoring equations were scaled and included. Lastly, the usefulness of the algorithm was demonstrated by comparing the *quality-quotient* or *qq* to the quality estimations of three individual trained in the use of microarrays.

The quality information provided by the *qq* can be applied to analyze variation between microarray replicates. The strength of the score pertaining to the spots on an individual chip could be used as a weighting factor. The weighted average of a particular feature from the set of replicates would most prominently represent the feature values from the strongest *qq* scores, and thus the best quality features.

Future experiments could include comparing the results of the *qq* function to the results of similar, image-based quality programs. It might also be useful to expand the number of "experts" to reduce variation in comparison scores. Lastly, it seems that better squashing function can reduce *qq* variability considerably.

APPENDIX A

'R' CODE FOR AUTOMATION OF FUNCTIONS

The following is the 'R' code written in order to automate the microarray normalization procedure.

```
> readin  #function to read in the chip
function (x)
{
  a <- paste(x, ".txt", sep="")
  data <- read.genepix(a, sep="\t", header=T, skip=0)
  data
}

> chipdef  #function to define chip parameters
function() {
  ngrid.r <- 8; ngrid.c <- 4
  nspot.r <- 12; nspot.c <- 14
 list(nspot.r = as.integer(nspot.r), nspot.c = as.integer(nspot.c),
      ngrid.r = as.integer(ngrid.r), ngrid.c = as.integer(ngrid.c))
}

> normal  #function that performs normalization on chips
function()
{
  library(sma)

  cat("What is the chip name?: ")
  prefix <- readline()

  #cat("How many chips are there?: ")
  #quantity <- readline()
  #quantity <- as.integer(quantity)
  #namelist <- paste(prefix, 1:quantity, sep="");   #constructs a list of prefix and number

  chip.setup <- chipdef()  #need done just once. chip dimension

  #for(i in 1:quantity)       #read in chip function; one cycle per chip
  #{
     x <- prefix
     chip <- readin(x)

    rm(x)
  cat("Done reading in chips.\n")
```

```
#**************************init.data function included here as list of commands

 name.G <- "Gmed"
 name.Gb <- "Gbmed"

name.R <- "Rmed"
name.Rb <- "Rbmed"

  res <- list(R = NULL, G = NULL, Rb = NULL, Gb = NULL)

      tmp <- eval(as.name("chip"))[, c(name.R, name.G, name.Rb, name.Gb)]

      res$R <- cbind(res$R, as.numeric(as.vector(tmp[, 1])))
      res$G <- cbind(res$G, as.numeric(as.vector(tmp[, 2])))
      res$Rb <- cbind(res$Rb, as.numeric(as.vector(tmp[, 3])))
      res$Gb <- cbind(res$Gb, as.numeric(as.vector(tmp[, 4])))

  chip.data <- res

  cat("Finished creating the dataset.\n")

  rm(name.G,name.Gb,name.R,name.Rb, tmp, res)

#**************************************************end of function init.data

  chip.lratio <- stat.ma(chip.data, chip.setup, norm="p") #calculate lowess

  filename <- paste(prefix,"lratio.txt", sep="")
  filename <- as.character(filename)

 write.table(chip.lratio, file=filename, sep="\t", row.names=F)
  #}
}
```

APPENDIX B

THE QUALITY-QUOTIENT FORMULAS

The formulas exist in a separate spreadsheet. They are copied as a whole and placed in cell 'AR' after cell 'AQ', the last cell of the Excel results worksheet ("flags"). Only the first row of formulas need be copied. The rest can be obtained by copying the first row down. The formula for each cell of the quality-quotient is given below.

1. Cell AR (635nm feature divided by 635nm background)
   Column heading: "F/B 635"
   Formula: "=I3/L3"

2. Cell AS (635 Feature-Background ratio scoring function)
   Column heading: "F/B 635 score"
   Formula: "=(2/PI())*ATAN(2*(AR3-1.4))"

3. Cell AT (635 feature divided by 635 background)
   Column heading: "CV 635"
   Formula: "=K6/J6"

4. Cell AU (635 CV scoring function)
   Column heading: "CV 635 score"
   Formula: "=1-(1.2/(1.2+0.7*(0.1/AT6^5)))"

5. Cell AV (532nm feature divided by 532nm background)
   Column heading: "F/B 532"
   Formula: "=R6/U6"

6. Cell AW (532 Feature-Background ratio scoring function)
   Column heading: "F/B 532 score"
   Formula: "=(2/PI())*ATAN(2*(AV6-1.4))"

7. Cell AX (532 feature divided by 532 background)
   Column heading: "CV 532"
   Formula: "=T6/S6"

8. Cell AY (532 CV scoring function)
   Column heading: "CV 532 score"
   Formula: "=1-(1.2/(1.2+0.7*(0.1/AX6^5)))"

9. Cell AZ
   Column heading: "Final scores>"
   Formula: none (leave blank)

10. Cell BA (Only counts red (635) FB scores above 0)
    Column heading: "Red F/B score"
    Formula: "=IF(AS6<0,0,AS6)"

11. Cell BB (Only counts green (532) FB scores above 0)
    Column heading: "Green F/B score"
    Formula: "=IF(AS6<0,0,AS6)"

12. Cell BC (organize scores by copying red CV score)
    Column heading: "Red CV score"
    Formula: "=AU6"

13. Cell BD (organize scores by copying green CV score)
    Column heading: "Green CV score"
    Formula: "=AY6"

14. Cell BE (Average of the three Ratio Quantities)
    Column heading: "Mean Ratio"
    Formula: "=AVERAGE(AA6,AC6,AF6)"

15. Cell BF (Mean difference of the three ratio quantities from average)
    Column heading: "MD"
    Formula: "=(ABS(AA6-BE6)+ABS(AC6-BE6)+ABS(AF6-BE6))/3"

16. Cell BG (Mean difference divided by average)
    Column heading: "PD"
    Formula: "=BF6/BE6"

17. Cell BH (Score for Percent difference)
    Column heading: "PD score"
    Formula: "=IF(BG6>1,0,1-(2.1/(2.1+0.6*(0.2/BG6^3.1))))"

18. Cell BI (Average CV scores for both wavelengths)
    Column heading: "CVcomp"
    Formula: "=(BC6+BD6)/2"

19. Cell BJ (Average FB scores for both wavelengths)
    Column heading: "F/Bcomp"
    Formula: "=(BC6+BD6)/2"

20. Cell BK (Final score-average of the FBcomp, CVcomp, and PD)
    Column heading: "QQ"
    Formula: "=AVERAGE(BH6:BJ6)

APPENDIX C

EXPERT SCORES

This chart shows the results: each expert score for spot 1-60 are listed followed by the expert average and QQ score.

		Experts			Expert		Modified	
#	Spot ID	DA	JE	JA	Average	QQ	QQ	Difference
1	A-10-1	3	3	5	3.67	0.97	4.86	1.19
2	A-10-2	5	4	4	4.33	0.55	2.75	1.58
3	A-10-3	3	3	3	3.00	0.63	3.13	0.13
4	A-10-4	3	3	3	3.00	0.55	2.77	0.23
5	A-10-5	3	3	2	2.67	0.64	3.19	0.52
6	A-10-6	4	4	2	3.33	0.99	4.94	1.61
7	A-10-7	3	3	4	3.33	0.25	1.27	2.07
8	A-10-8	3	3	3	3.00	0.63	3.13	0.13
9	A-10-9	3	4	2	3.00	0.72	3.60	0.60
10	A-10-10	5	4	2	3.67	0.56	2.82	0.85
11	A-10-11	3	4	2	3.00	0.67	3.33	0.33
12	A-10-12	5	5	3	4.33	0.35	1.74	2.59
13	A-5-1	4	3	3	3.33	0.60	3.00	0.33
14	A-5-2	5	3	4	4.00	0.90	4.48	0.48
15	A-5-3	4	3	3	3.33	0.71	3.55	0.22
16	A-5-4	4	4	3	3.67	0.52	2.60	1.07
17	A-5-5	3	3	4	3.33	0.83	4.17	0.84
18	A-5-6	5	5	5	5.00	1.00	5.00	0.00
19	A-5-7	4	3	3	3.33	0.50	2.52	0.81
20	A-5-8	3	3	2	2.67	0.68	3.39	0.73
21	A-5-9	4	3	3	3.33	0.79	3.97	0.64
22	A-5-10	3	4	3	3.33	0.63	3.16	0.17
23	A-5-11	2	3	2	2.33	0.54	2.71	0.38
24	A-5-12	1	1	1	1.00	0.29	1.46	0.46
25	A-6-1	4	3	3	3.33	0.82	4.10	0.77
26	A-6-2	3	3	3	3.00	0.72	3.58	0.58
27	A-6-3	4	3	2	3.00	0.52	2.58	0.42
28	A-6-4	5	3	5	4.33	0.92	4.58	0.24
29	A-6-5	4	3	3	3.33	0.50	2.50	0.83
30	A-6-6	3	3	4	3.33	0.59	2.93	0.40
31	A-6-7	4	3	3	3.33	0.45	2.24	1.09
32	A-6-8	4	3	3	3.33	0.61	3.07	0.26
33	A-6-9	3	4	3	3.33	0.69	3.47	0.14
34	A-6-10	4	3	2	3.00	0.51	2.57	0.43

#	Spot ID	Experts			Expert Average	QQ	Modified QQ	Difference
		DA	JE	JA				
35	A-6-11	4	4	2	3.33	0.30	1.50	1.83
36	A-6-12	1	1	1	1.00	0.50	2.51	1.51
37	B-4-1	4	4	3	3.67	0.69	3.47	0.20
38	B-4-2	3	4	3	3.33	0.74	3.70	0.36
39	B-4-3	5	5	5	5.00	0.99	4.95	0.05
40	B-4-4	5	3	4	4.00	0.91	4.55	0.55
41	B-4-5	3	3	4	3.33	0.87	4.33	1.00
42	B-4-6	4	3	5	4.00	0.96	4.79	0.79
43	B-4-7	3	3	3	3.00	0.72	3.62	0.62
44	B-4-8	3	4	2	3.00	0.69	3.44	0.44
45	B-4-9	3	4	2	3.00	0.62	3.12	0.12
46	B-4-10	4	4	4	4.00	0.94	4.69	0.69
47	B-4-11	2	3	4	3.00	0.84	4.22	1.22
48	B-4-12	5	5	5	5.00	0.97	4.87	0.13
49	C-5-1	4	4	3	3.67	0.55	2.75	0.91
50	C-5-2	4	4	3	3.67	0.70	3.52	0.14
51	C-5-3	3	3	4	3.33	0.88	4.41	1.08
52	C-5-4	3	3	3	3.00	0.79	3.97	0.97
53	C-5-5	4	3	3	3.33	0.47	2.35	0.98
54	C-5-6	3	4	3	3.33	0.73	3.63	0.29
55	C-5-7	2	3	2	2.33	0.58	2.90	0.57
56	C-5-8	4	3	3	3.33	0.74	3.72	0.39
57	C-5-9	1	2	1	1.33	0.56	2.80	1.47
58	C-5-10	1	3	2	2.00	0.73	3.66	1.66
59	C-5-11	2	3	2	2.33	0.77	3.86	1.53
60	C-5-12	5	3	4	4.00	0.79	3.95	0.05

REFERENCES

1. Duggan D.J., Bittner M., Chen Y., Meltzer P., Trent J.M. Expression Profiling Using cDNA Microarrays. *Nature Genetics*. **21**, 10-14 (1999).

2. Schena M., Shalon D., Davis R.W., Brown P.O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*. **270**, 467-470 (1995).

3. Lipshutz R.J., Fodor S.P.A, Gingeras, T.R, Lockhart D.J. High Density Synthetic Oligonucleotide Arrays. *Nature Genetics*. **21**, 20-24 (1999).

4. Deyholos M.K., Galbraith D.W., High Density Microarrays for Gene Expression Analysis. *Cytometry*. **43**, 229-238 (2001).

5. Hoheisal J.D., Vingron M. Transcription Profiling: is it Worth the Money? *Research Microbiology*. **151**, 113-119 (2000).

6. Kahn D.E., Lam J.S., Carmel J.B., Recce M., Soteropoulos P., Tolias P., Hart R.P. Assessing Hybridization Conditions with Rat Oligonucleotide Microarrays: Automation Improves Reproducibility. Submitted to *Nucleic Acids Research*, March 2002.

7. Handran S., Zhai J.Y., *Biological Relevance of Genepix Results*. Genepix Manual. Axon Instruments Inc. www.axon.com.

8. Long A.D., Mangalam H.J., Chan B.Y.P., Tolleri L., Hatfield G.W., Baldi P. Improved Statistical Inference from DNA Microarray Data Using Analysis of Variance and a Bayesian Statistical Framework. *The Journal of Biological Chemistry*. **276**, 19937-19944 (2001).

9. Motulsky, H., *Intuitive Biostatistics*. Oxford University Press, 1995. (207-209).

10. Spiegel M.R., Stephens L.J., *Theory and Problems of Statistics*, MrGraw Hill, 1998. (241-245).

11. Beibarth T., Fellenberg K., Brors B., Arribas-Prat R., Boer J.M., Hauser N.C., Scheideler M., Hoheisel J.D, Schutz G., Poustka A., Vingron M. Processing and Quality Control of DNA Array Hybridization Data. *Bioinformatics*. **16**, 1014-1022 (2000).

12. Samartzidou H., Turner L., Houts T. Lucidea Microarray Scorecard. An Integrated Tool for Validation of Microarray Gene Expression Experiments. *Life Science News, Amerisham Pharmacia Biotech*. (2001)

13. Tseng G.C., Oh M., Rohlin L., Liao J.C., Wong W.H. Issues in cDNA Microarray Analysis: Quality Filtering, Channel Normalization, Models of Variations, and Assessment of Gene Effects. *Nucleic Acids Research.* **29**, 2549-2557 (2001).

14. Brown C.S., Goodwin P.C., Sorger P.K. Image Metrics in the Statistical Analysis of DNA Microarray Data. *PNAS.* **98**, 8944-8949 (2001).

15. Wang X., Ghosh S., Guo S., Quantitative Quality Control in Microarray Image Processing and Data Acquisition. *Nucleic Acids Research.* **29**, (2001).

16. Rajeevan M.S., Ranamukhaarachchi D.G., Vernon S.D., Unger E.R. Use of Real-Time Quantitative PCR to Validate the Results of cDNA Array and Differential Display PCR Technologies. *Methods* **25**, 443-451 (2001).

17. Burgess, J.K. Gene Expression Studies Using Microarrays. *Clinical and Experimental Pharmacology and Physiology.* **28**, 321-328 (2001).

18. Bouchez D., Höfte H., Functional Genomics in Plants. *Plant Physiology.* **118**, 725-732 (1998).

19. Ihaka R., Gentleman R., R: A Language for Data Analysis and Graphics. *Journal of Computer Graphics and Statistics.* **5,** 229-314 (1996).

20. Dudoit S., Yang Y.H., Statistical Microarray Analysis. *The SMA Package.* http://www.stat.berkeley.edu/users/terry/zarray/Html/smacode.html. (2002).

21. Stears R.L., Getts R.C., Gullans S.R. A Novel, Sensitive Detection System for High-Density Microarrays Using Dendrimer Technology. *Physiological Genomics.* **3**, 93-99 (2000).

22. Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., Speed T. P. Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation. *Nucleic Acids Research.* **30**, (2002).

23. Carmel J.B., Galante A., Soteropoulos P., Tolias P., Recce M., Young W., Hart R.P. Gene Expression Profiling of Acute Spinal Cord Injury Reveals Spreading Inflammatory Signals and Neuron Loss. *Physiological Genomics.* **7**, 201-213 (2001).

24. Nesic O., Svrakic N.M., Xu G-Y., McAdoo D., Westlund K.N., Hulsebosch C.E., Zeiming Ye, Galante A., Soteropoulos P., Tolias P., Young W., Hart R.P., Perez-Polo J.R. DNA Microarray Analysis of the Contused Spinal Cord: Effect of the NMDA Receptor Inhibition. *Journal of Neuroscience Research.* **68**, 406-423 (2002).